AWARD NUMBER: W81XWH-14-1-0080

TITLE: Total RNA Sequencing Analysis of DCIS Progressing to Invasive Breast Cancer.

PRINCIPAL INVESTIGATOR: Christopher B. Umbricht, MD, PhD

CONTRACTING ORGANIZATION: Johns Hopkins University
Baltimore, MD 21205

REPORT DATE: September 2015

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE September 2015	**2. REPORT TYPE** Annual	**3. DATES COVERED** 1 Sep 2014 - 31 Aug 2015

4. TITLE AND SUBTITLE Total RNA Sequencing Analysis of DCIS Progressing to Invasive Breast Cancer.	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER W81XWH-14-1-0080
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Christopher B. Umbricht, MD, PhD	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
E-Mail: cumbrich@jhmi.edu	**5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Johns Hopkins University Baltimore, MD 21205	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT: This project is designed to complement an ongoing a multi-institutional, NIH-funded study of genetic and epigenetic alterations of pre-invasive DCIS and no previous or concurrent invasive breast cancer (IBC) that either progressed to IBC (cases) or had no further breast cancer events (controls), with an in-depth analysis of expression data on the entire range of informative RNA categories, including mRNA, miRNA, and lncRNA, as well as their splice variants. During the current reporting period, we have completed the accrual, initial characterization and processing of samples from 5 collaborating institutions. Among 229 patient tissues processed for the discovery phase, a total of 196 samples, 98 cases and controls, passed Q/C for both DNA and RNA extracts. All samples have undergone a comprehensive DNA methylome analysis using the Illumina 450K CpG arrays, with excellent call rates, the bioinformatics analysis is ongoing. Because of the high failure rate in generating high quality libraries for RNA Sequencing from our limited samples, we proceeded with a pilot study of the newly released Affymetrix HTA 2.0 arrays, with excellent results, and have begun our comprehensive transcriptome analysis using this platform. We have also initiated a collaboration with Dr. C. Perou at UNC to maximize the possibility of a successful RNA Sequencing effort.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	Unclassified		**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Page

1. Introduction.

Our overall goal is to develop predictive markers that will be useful in identifying the minority of cases of preinvasive breast cancer (DCIS), that do in fact progress to invasive disease (IBC). This project is designed to complement an ongoing a multi-institutional, NIH-funded study of genetic and epigenetic alterations of pre-invasive DCIS and no previous or concurrent invasive breast cancer (IBC) that either progressed to IBC (cases) or had no further breast cancer events (controls), with an in-depth analysis of expression data on the entire range of informative RNA categories, including mRNA, miRNA, and lncRNA, as well as their splice variants.

The SPECIFIC AIMS, which are substantially unchanged, are as follows:

Specific Aim 1: Perform a comprehensive analysis of the DCIS transcriptomes of a multicenter cohort of patients with either progression to invasive breast cancer, or with over 10 years of disease free survival. The objective is to obtain a comprehensive catalog of transcriptome alterations in DCIS, covering differential transcription levels, alternate splicing variation, and non-coding RNA expression (both miRNA and lncRNA), using a state-of-the-art platform.

Specific Aim 2: Perform bioinformatic analyses identifying signatures that are specific for high-risk DCIS, and integrate the sequencing data with complementary datasets from the same cohort. The objective is to select small sets of features that together discriminate classes, while avoiding over-fitting and benefiting from cross-platform validation. We will assemble small, non-overlapping models for validation in subsequent aims.

Specific Aim 3: Develop a panel of multiplex assays that can be used in minimal routine clinical material to predict long-term outcome in DCIS, and optimize performance on in-house DCIS samples. Candidate marker sets will be characterized biochemically and marker-specific assays applicable to high throughput analysis of clinical samples will be developed. Markers that perform well will be combined into multiplex quantitative PCR and Nanostring assays that can be tested for optimal prognostic performance on in house tissue samples.

Specific Aim 4: Validate the results in independent, population-based test cohorts of DCIS patients with progressive disease vs. DCIS patients without recurrent disease, using the newly developed assays. Our objective is to prospectively test our DCIS assay on an independent test set in order to obtain a realistic assessment of its potential positive and negative predictive power.

2. Keywords

Preinvasive breast cancer (DCIS); Transcriptome; Prognostic markers; splice variant analysis; non-coding RNA; formalin-fixed paraffin-embedded (FFPE) tissue.

3. Accomplishments

During the current reporting period, we have completed the accrual, initial characterization and processing of samples from 5 collaborating institutions, with particular focus on yields sufficient for various planned array platforms, given the often small size of the DCIS lesions. We have now completed the necessary DNA and RNA preparations for the initial discovery phase, with sufficient samples passing our Q/C testing to include 98 cases (progression to invasive breast cancer) and 98 controls (no further breast cancer or DCIS).

DNA/RNA co-extraction:

Areas with DCIS were annotated on H&E slides by a pathologist. Using that H&E slide as a guide, neighboring unstained FFPE slides were macrodissected using a sterilized blade to enrich for at least 70% tumor. DNA and RNA were co-extracted from the macrodissected tissue using the Qiagen Allprep DNA/RNA FFPE Kit (Qiagen, Germany) with a modified protocol optimized for extracting nucleic acids from FFPE material.

DNA and RNA were quantified using dsDNA and RNA Broad Range Assay Kits respectively on the Qubit 2.0 Fluorometer (Life Technologies). The BioRad Experion Automated Electrophoresis System RNA kit was used to analyze the quality of the RNA samples. RNA that were completely degraded were re-extracted where tissue was available.

These results are summarized in the following figures:

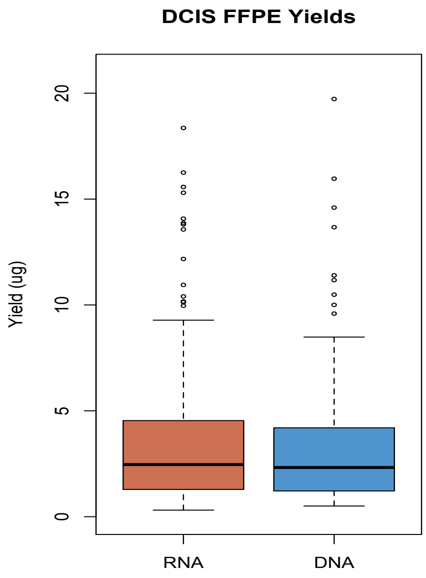


Figure 1. Yields of 196 discovery samples.

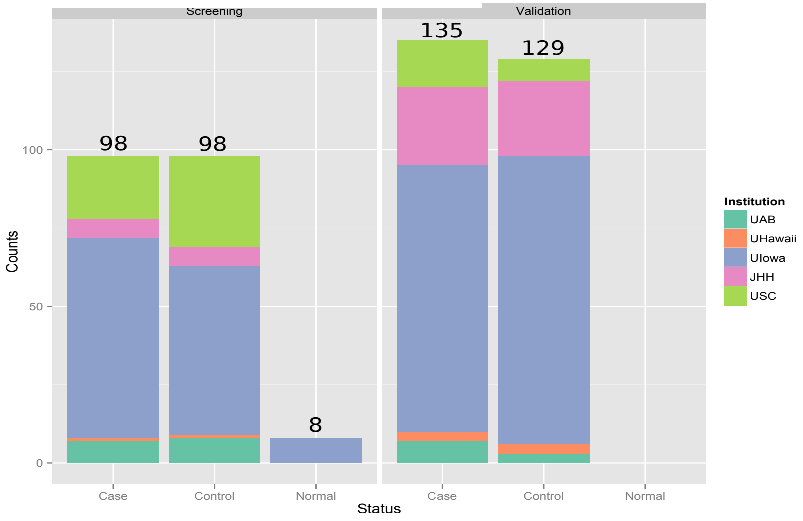


Figure 2: Sample distribution for the study.

DCIS Sample sets

Among 229 patient tissues processed for the discovery phase, a total of 204 samples from 196 patients were analyzed using the 450K microarray. The 25 samples that either failed QC or had insufficient material after bisulfite conversion and will be reserved for the validation phase.

A total of 264 samples are available for validation phase of the study.

Table 1: Sample distribution

	JHH	UAB	UHawaii	UIowa	USC
Discovery Phase					
Case	6	7	1	64	20
Control	6	8	1	54	29
Normal	0	0	0	8	0
Total	**12**	**15**	**2**	**126**	**49**
Validation Phase					
Case	25	7	3	85	15
Control	24	3	3	92	7
Total	**49**	**10**	**6**	**177**	**22**

We have completed the bisulfite conversion of DNA samples for the methylome analysis, and have just completed the methylome microarray chip assays with excellent technical call rates. The bioinformatic analysis of the methylome is ongoing.

In light of the often limiting amounts of nucleic acids we can obtain from our archival tissue samples, particularly given the often small DCIS lesion sizes, we have also proceeded with our development and investigation of a computational approach we have called EPICOPY to obtain reliable copy number variation (CNV) data from the methylome array data, thereby decreasing the DNA requirements in half.

Methylome Analysis (Illumina Human Methylation 450K Methylation Microarray):

DNA was bisulfite treated using a modified protocol, per Appendix 1 of manufacturer recommendations, of the Zymogen EZ DNA Methylation Kit. Bisulfite-treated DNA were restored using Illumina's DNA Restoration Kit and processed for the Illumina Human Methylation 450K array per manufacturer instructions. Raw Illumina .idat files were read and analyzed using the minfi package in the R statistical environment.

Samples were assessed for good performance on the array using detection p-values, a metric implemented by Illumina to identify probes detected with confidence. Samples less than 90% of probes detected were removed from the analysis and probes undetected in any of the samples were filtered away.

The analysis workflow is highlighted in Figure 3. Briefly, samples were assessed for chip performance through the detection p-value metric, which is a measure of confidence of signal intensities. Samples with call rates (probes detected) of less than 95% were filtered away and probes that were not called in at least a single sample were removed from further analysis.

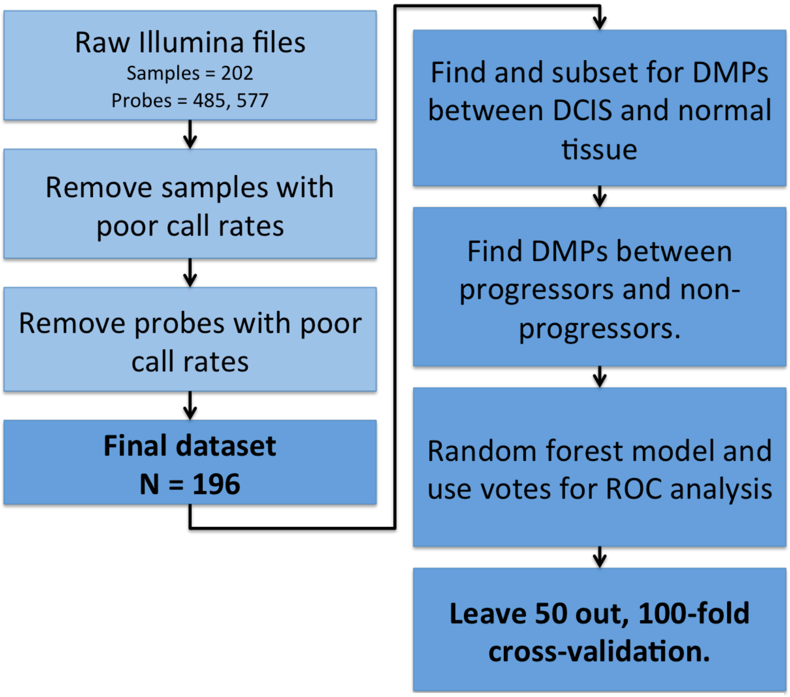

Raw Illumina files
Samples = 202
Probes = 485, 577
Remove samples with poor call rates
Remove probes with poor call rates
Final dataset
N = 196
Find and subset for DMPs between DCIS and normal tissue
Find DMPs between progressors and non-progressors.
Random forest model and use votes for ROC analysis
Leave 50 out, 100-fold cross-validation.

Figure 3: Methylation analysis workflow

Linear models for microarray analysis (limma) was used to identify differentially methylated probes (DMPs) between DCIS and normal breast tissue. Sub-setting for these DCIS-specific probes, limma was used again to identify DMPs between progressors and non-progressors. Absolute t-statistics were used to choose between 1 – 100 top probes that best distinguish progressors from non-progressors. This subset of probes was then used in a random forest model with 10,000 trees to build a model to best distinguish non-progressors from progressors. An ROC analysis was performed with the votes for progressors from the random forest model as the predictor to estimate probe set performance. The specificity at 95% sensitivity is used as the metric for assessing probe set performance. Following that, a 100-fold, leave-50-out cross-validation experiment was performed to assess this method of selecting probe sets.

181 DCIS and 13 normal tissue samples passed QC and were used for the analysis. Currently, little molecular profiling, be it IHC or FISH, drives clinical decision. To emulate what happens in the clinical setting, we identified DMPs from cases and controls naïve of common breast cancer markers such as ER/PR/HER2.

ROC analysis of the resulting random forest model reveals an AUC of 0.766 and a specificity of 34.4% at 95% sensitivity (Table 2). Furthermore, leave-50-out 100-fold cross-validation reveals a mean specificity of 12.5% at 95% specificity.

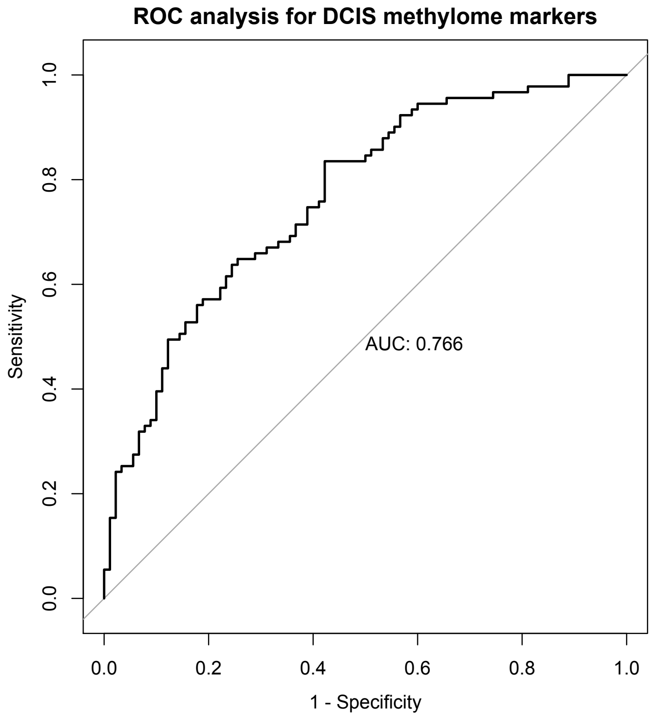


Figure 4: ROC analysis for DCIS marker set

Table 2: Sensitivity and specificity of DCIS marker set for distinguishing progression.

Sensitivity	Specificity
0.95	**0.11**
0.98	**0.11**
1.00	**0.34**

Copy number variation from methylation data:

We have developed a computational method, Epicopy, to obtain copy number information from methylation microarrays. Using Epicopy, we obtained segmented copy number information for these DCIS samples and performed GISTIC 2.0 analysis to identify regions of recurrent CNVs across all samples.

The copy number profiles of the same samples that passed QC were estimated using Epicopy and analyzed using GISTIC (Figure 5).

Of note, we observe a loss of the 17p arm and gain of 17q arm, which have been report by previous studies to be hallmarks of non-progressive and progressive disease respectively.

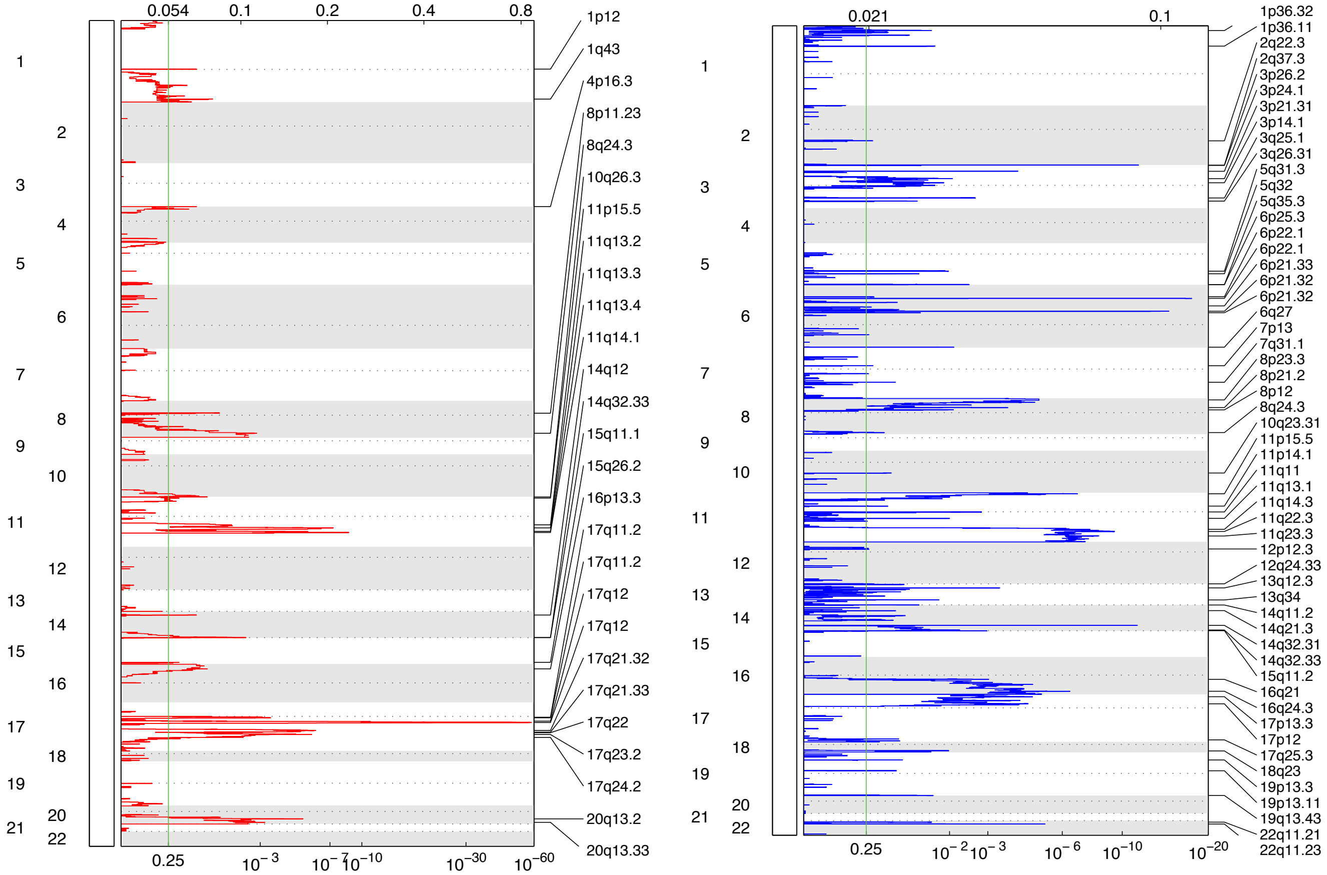


Figure 5 GISTIC results for DCIS copy number profile.

Future analyses:

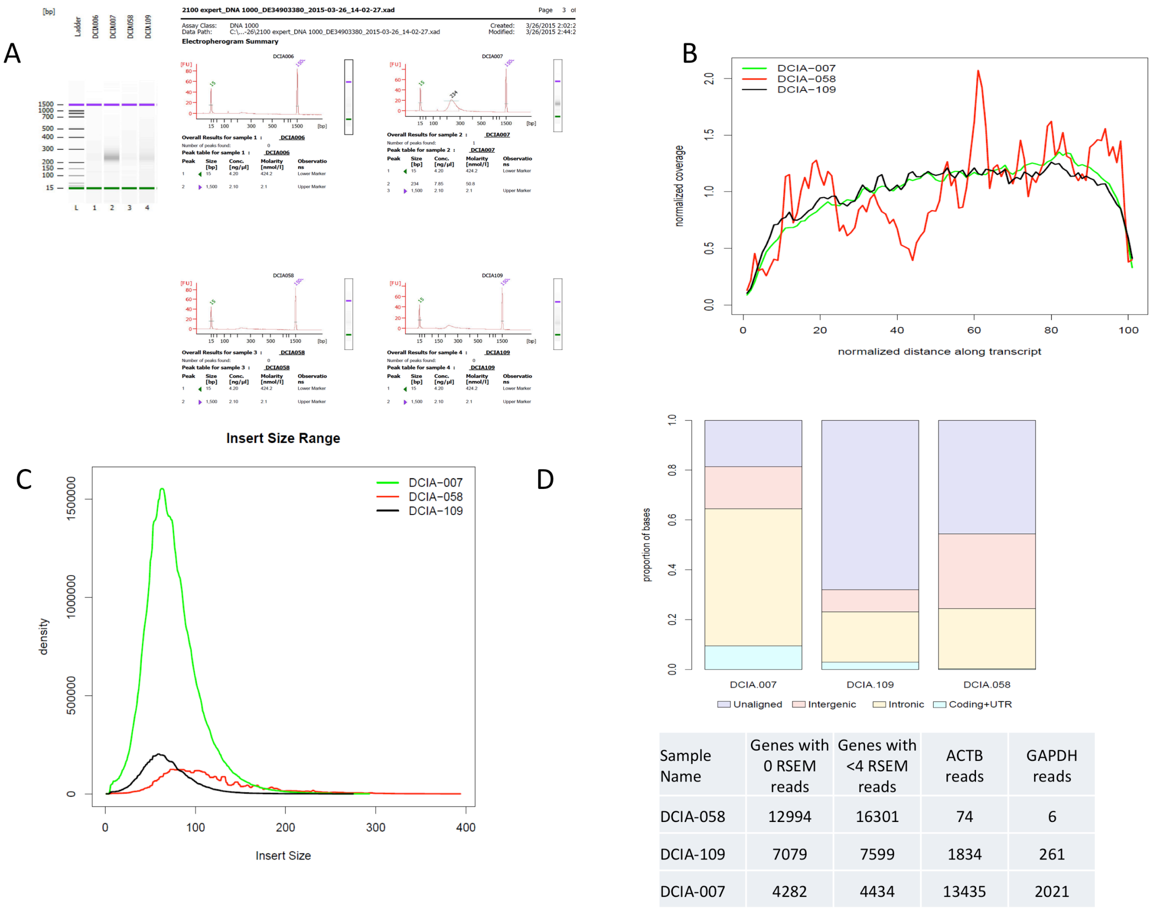
We expect to obtain transcriptome data for these samples by the end of the year, which will allow us to identify molecular phenotypes, such as PAM50 and ER/HER2 status, which will allow us to classify these samples into more appropriate molecular and biological groups. Fisher's Exact test will be performed to identify enrichment of different subtypes in either progressors or non-progressors.

A two-group analysis will be performed on the copy number information, using both GISTIC and Fisher's Exact test to identify regions that are copy number altered in either progressors or non-progressors.

A risk model for progression will be trained from a combination of all three molecular profiles.

Transcriptome pilot experiments

Initial test experiments using our institutional core facility to determine the quality of data obtained from our very challenging samples (due to the combined consequences of very small lesions (often <5mm) and the deleterious effects of formalin fixation and long term storage of archival samples, unavoidable because of the rarity of DCIS progressing to IBC) were unsatisfactory. We therefore established a collaboration with Dr. Charles Perou at the University of North Carolina, one of the pioneers of gene expression analysis in breast cancer. We sent RNA extracts of 4 representative DCIS samples, 2 cases and 2 controls, for an initial test of their established RNA-Seq analysis pipeline.



Sample Name	Genes with 0 RSEM reads	Genes with <4 RSEM reads	ACTB reads	GAPDH reads
DCIA-058	12994	16301	74	6
DCIA-109	7079	7599	1834	261
DCIA-007	4282	4434	13435	2021

Figure 6. RNA-Seq of 4 DCIS samples.

As summarized in Figure 6, only 3 of the four samples yielded a library, and only 2 of the samples resulted in any aligned reads of coding transcripts. In the estimation of the experts at UNC, only one sample, DCIA-007, produced a result that could be used for further analysis. In light of these results, we opted to investigate the newly released HTA2.0 array from Affymetrix. This array contains a combination of probes detecting both coding and non-coding transcripts, as well as so-called junction probes covering exon-intron boundaries, enabling a detailed analysis of alternative splicing, one of our original goals in this project.

An additional advantage of this approach is that the RNA requirements for this analysis are in the 10-20 ng range, even for poor quality RNA samples. Therefore, we can perform

this analysis without jeopardizing a possible later RNA-Seq analysis if the technical hurdles currently preventing that can be overcome.
Our initial pilot on the HTA2.0 array consisted of the same 4 samples used for the RNASeq pilot, tested at 1,10, and 20 ng input RNA levels (the recommended amount is 10ng). All 4 samples produced % present call rates ranging from 37-40, and even the sample that had completed failed to generate a RNASeq library showed 31% present calls. These rates are close to the ones achieved with high quality RNA from frozen tissue or call line-derived RNA.

DCIS sample#	#006	#007	#058	#109
All probeset % called	37	40	31	39
Positive control % called	58	41	42	56
% of control called	65	97	73	68

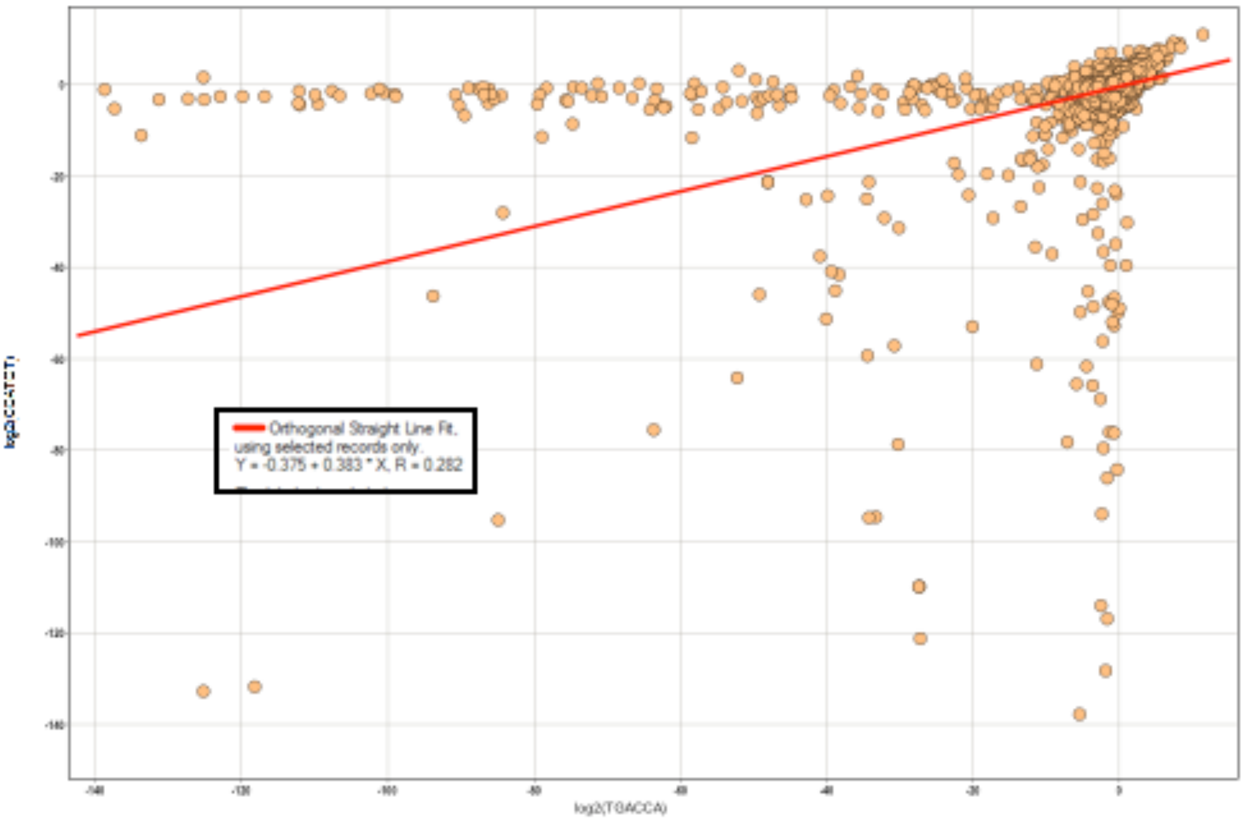


Figure 7. Correlation of mapped RNA-Seq transcripts of DCIS#007 with HTA2.0 array.

Figure 7 illustrates the correlation achieved between the best RNASeq result and the corresponding HTA2.0 array, indicating that at higher levels of expression, the two platforms produced consistent results, at least for the sample that had interpretable RNASeq data.

Based on these results, we are proceeding with the HTA2.0 array analysis of our DCIS discovery cohort, while continuing our attempts to improve the RNASeq analysis of poor quality RNA in collaboration with Dr. Perou's group at UNC.

4. Impact

N/A

5. Changes/Problems

See discussion of our results in section 3. After failing to obtain acceptable results from our initial test samples using RNASeq, we successfully piloted the new Affymetrix HTA2.0 arrays using limited amounts of RNA extracted from our DCIS cohort.

6. Products

N/A

7. Participants & Other Collaborating Organizations

Charles M. Perou, Ph.D
The May Goldman Shaw Distinguished Professor of Molecular Oncology Departments of Genetics, and Pathology & Laboratory Medicine
Lineberger Comprehensive Cancer Center Marsico Hall, 5th floor, CB#7295
125 Mason Farm Road
The University of North Carolina at Chapel Hil Chapel Hill, NC 27599

8. Special Reporting Requirements

N/A

9. Appendices

N/A